SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

Ak. god. 2018/2019.

Danijel Blazsetin

Analiza distribucije glagolskih vrsta u općem korpusu hrvatskoga jezika

Završni rad

Mentor: dr. sc. Petra Bago, doc.

Zagreb, rujan 2019.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Sadržaj

1. Uvod

Ovaj se rad bavi frekvencijskom distribucijom glagolskih vrsta. Statistički su podaci izrađeni na jezičnom uzorku izlučenom iz mrežnog korpusa *hrWaC*. U slavenskim jezicima klasificiranje glagola u glagolske vrste ima dugu tradiciju. Temelj je glagolskih klasifikacija morfemska raščlamba glagola, stoga će se istaknuti najvažniji pojmovi vezani za morfologiju glagola. U radu će se donijeti kratki pregled bavljenja klasificiranjem glagola u vrste i ukazati na trendove u suvremenome hrvatskom jezikoslovlju. Opisat će se prema kojoj je glagolskoj klasifikaciji izrađen model i koji su uzroci toga odabira. U hrvatskim se preskriptivnim priručnicima ne mogu naći pouzdani i egzaktni podaci o frekvenciji glagolskih vrsti, a može se reći da se u hrvatskoj filologiji, izuzev svega nekoliko radova, rijetko problematizira frekvencija, plodnost i prototipnost glagolskih vrsta. Ulaženjem računarstva u lingvistiku ili lingivstike u računarstvo nastalo je novo, interdisciplinarno znanstveno područje koje se naziva računalnom lingvistikom ili obradom prirodnog jezika, ovisno o pristupu problematici. Ovakav pristup omogućava obradu ogromne količine podataka u vrlo malo vremena, bez velike količine ljudskih resursa. Koristeći metode obrade prirodnog jezika, ovaj rad nudi model za automatsku glagolsku klasifikaciju koji se demonstrira na jezičnom uzorku iz korpusa. Za potrebe ovoga rada bilo je nužno izraditi reprezentativni jezični uzorak koji se sastoji od triju skupina: najfrekventniji glagoli, srednjofrekventni glagoli i rijetki glagoli. Izrađena se frekvencijska distribucija glagola u vrste uspoređuje s dvjema hrvatskim gramatikama i dvjema opsežnijim statistikama te se ukazuje na podudarnosti, razilaženja i nelogičnosti koje se javljaju pri komparativnoj analizi podataka. U radu se uz usporedbu raznih statistika opisuju prednosti i mane pri analizi mrežnog korpusa te se ističu problemi proučavanoga jezičnoga uzorka. Opisuju se najčešće pogreške koje se mogu kategorizirati, a koje se dosad nisu sustavno opisivale. Izrađen računalni program ne opisuje se detaljno, ali rad ističe njegove najvažnije stavke i objašnjava njihovu funkciju.

Ključne riječi: *korpusna lingvistika, obrada prirodnoga jezika, glagolska klasifikacija, hrvatski jezik, hrvatske gramatike*

2. Glagoli i glagolske vrste

Priđe li se temi ovoga rada bilo prvenstveno iz perspektive obrade prirodnog jezika bilo iz perspektive lingvistike mora se odgovoriti na pitanje koje se uporno nameće, a zapravo otvara prostor ogromnim studijama i raspravama: što je zapravo glagol? Svrha ovoga rada nije opis glagolske tipologije ili prikaz rasprave o definiciji glagola u hrvatskim i stranim preskriptivnim i deskriptivnim priručnicima, stoga će se o glagolu govoriti samo onoliko koliko je nužno za shvaćanje predmeta.[1] Prema Barić *et al.* (1995: 222) glagoli su „promjenljive riječi kojima se izriču procesi – radnja, stanje i zbivanje. Karakteriziraju ih kategorije vida, lica, načina, vremena i stanja". Ovakva je definicija univerzalna, odnosno glagolima se u jezicima smatraju riječi koje izriču kakvu radnju (Marković 2012: 179).

Ono što je od same definicije glagola puno važnije za rad su glagolske vrste. Glagolske (konjugacijske) su vrste inherentna glagolska kategorija.[2] Riječ je o vrsti jezične ekonomije prema kojoj se pri fleksiji glagoli okupljaju u skupine prema svojim fleksijskim i derivacijskim morfovima (*ibid*: 197).[3] Na sljedećem se primjeru vidi koji morf (koji dio) glagola izražava konjugacijsku vrstu glagola (zasad zanemarimo imena glagolskih vrsta).

	IV. VRSTA	V. VRSTA 1. RAZRED
INF	rad-**i**-ti	gled-**a**-ti
Pz	rad-**i**-m	gled-**a**-m

[1] Za detaljni pregled glagola, glagolske tipologije, opis glagolskih kategorija u hrvatskom jeziku usp. Marković (2012: 179-229).

[2] Inherentne su glagolske kategorije one koje specificiraju, pobliže određuju i karakteriziraju predikaciju (glagol) (Marković 2012: 183).

[3] Morfem je „najmanji odsječak riječi kojemu je pridružen kakav sadržaj, tj koji ima kakvo značenje. (...) Izraz morfema zove se morf" (Barić *et al*: 1995: 96). Na neapstraktnoj jezičnoj razini uvijek govorimo o morfovima. Razlika između fleksijskih i derivacijskih morfema je ta što prvi služe za proizvodnju oblika riječi istoga leksema, dok potonji za proizvodnju oblika riječi novih leksema (Marković 2012: 52)

2.1. Različiti pristupi podjeli glagola u vrste

Glagoli se u slavenskim jezicima, pa tako i u hrvatskom, raspoređuju u glagolske vrste. U tradiciji su se ustalila dva glavna razvrstavanja, prema infinitivnoj (infinitivno-aoristnoj) i prema prezentskoj osnovi. Od dviju se glagolskih osnova tvore različiti oblici leksema. Infinitivna osnova služi za tvorbu infinitiva, aorista, glagolskog priloga prošlog i glagolskih pridjeva, dok prezentska služi za tvorbu prezenta, imperativa i glagolskoga priloga sadašnjeg. U hrvatskoj je gramatičkoj tradiciji zastupljenije razvrstanje prema infinitivnoj osnovi. Ova podjela slijedi model Josefa Dobrovskog koji glagole u staroslavenskom jeziku dijeli na šest glagolskih konjugacija ili vrsta.[4] Dobrovský definira sljedeće konjugacije: **I**. -ti se dodaje korijenu, **II**. infinitiv na -nǫ-ti, **III**. infinitiv na -ě-ti, **IV**. infinitiv na -i-ti, **V**. infinitiv na -a-ti, **VI**. infinitiv na -ova-ti. Mnoštvo je filologa slijedilo ovu podjelu, međutim, započevši s A. Schleicherom, niz se gramatika oslanja na podjelu prema prezentskoj osnovi. Stjepan Ivšić u svojoj Poredbenoj slavenskoj gramatici definira sljedeći model: **I**. prezentska osnova na -o-, -e-, **II**. prezentska osnova na -n-, -ne-, **III**. prezentska osnova na -jo- -je-, **IV**. prezentska osnova na -i-, **V**. čine atematski glagoli (usp. Marković 2012: 216-218, Bošnjak Botica 2013: 65-77).[5] Morfovi koji se javljaju u glagolima i koji određuju njihovu glagolsku vrstu nazivaju se tematskim morfom.[6] Kao što vidimo i u jednoj i u drugoj podjeli, pojavljuju se glagoli koji „imaju nešto, a to nešto je ništa“. To su glagoli s nultim tematskim morfom na koje se posebno obraća pozornost jer predstavljaju probleme pri računalnoj obradi glagola.

INF	pek-**ø**-ti	jed-**ø**-ti
Pz	pek-e-m	jed-e-m

U suvremenoj se hrvatskoj filologiji obično dva različita modela za razvrstavanje glagola oprimjeruju s dvjema *velikim i značajnim* gramatikama. Infinitivna se podjela oprimjeruje s gramatikom Eugenije Barić i njezinih suradnika (1995), a prezentska se podjela pokazuje na podjeli iz gramatike Silić i Pranjković (2005). Osim ovih podjela, naravno, postoji niz različitih modela koje su, među ostalim, ponudili Babić, Silić, Z. Babić (kasnije Jelaska), Tadić, Bošnjak Botica i Marković. Ovaj rad ne namjerava niti dati potpuni pregled svih modela ističući njihove prednosti i mane, niti komparirati razne podjele, niti ponuditi svoju

[4] Prema Markoviću (2012: 216) Dobrovský je podjelu mogao naći u slovačko-češkoj gramatici Slovaka P. Doležala.

[5] Atematskim se glagolima nazivaju glagoli bez tematskog sufiksa (*greb-ø-ti, grepsti*)

[6] Šojat, Srebačić i Štefanec (2013: 90) drugog su mišljenja i kažu: „ako oni *(tematski sufiksi, samoglasnici)* obilježavaju glagol kao glagol, drugim riječima, ako je njihov sadržaj „glagol” i govore o glagolskoj vrsti, preciznije bi bilo govoriti o sufiksima, a ne o tematskim samoglasnicima.”

podjelu. Tomu je više razloga. Prvotno i prije svega to nije zadaća rada s područja informacijskih znanosti. Međutim tomu ima i razloga koji su jezikoslovne naravi.

Prvo, kao što i Marković kaže, ponuditi savršenu podjelu nemoguće je (Marković 2012: 225-227). Pokušavamo li previše apstrahirati podjelu, izmiču nam iznimke, a nastojimo li popisati sve glagole koji se javljaju u jeziku gubimo preglednost sustava, njegovu apstraktnost. Drugo, čini se da bi se već o samoj komparaciji različitih podjela mogao napisati vrlo opsežan rad, a pokuša li ovaj rad dati detaljan opis barem najznačajnijih gramatika napisanih u posljednjim desetljećima, izmiče mu njegova glavna tema. Treće, da bi ovaj rad ponudio novu podjelu autor bi trebao imati uvid u čitavu povijest bavljenja glagolskim vrstama, odnosno glagolima općenito. Četvrto, u ovom je radu zapravo svejedno koja se podjela izabere kao temelj. Odabir može biti posve arbitraran. Manjim ili većim promjenama u programskom kodu možemo primijeniti različite podjele na jezični resurs i razmotriti *uspješnost* neke podjele. Peto, vrlo je teško, na temelju računalno obrađenog korpusa, bez ljudskog nadgledanja, zaključiti plodnost, ispravnost neke podjele. Numerički podaci o distribuciji po glagolskim vrstama u računalnom korpusu mogu nam pružati dobar (i zasad neviđen) uvid (iako on bio okvirni i iako njemu treba pristupiti s dozom opreznosti) u plodnost, prototipnost neke glagolske vrste. Podaci poput onih iznesenih u ovome radu rijetki su u hrvatskoj filologiji, odnosno „hrvatska filologija kronično ne pokazuje interes za matematičko, količinsko iskazivanje brojnosti pojedine fleksijske vrste. Ili smatra da to nisu važni podaci, što je jednako zabrinjavajuće“ (*ibid*: 219–220).

2.2. Podjela prema prezentskoj osnovi kao temelj frekvencijske analize

Radi proučavanja glagolske distribucije bilo je nužno izabrati podjelu koju će rad primijeniti na jezičnom resursu. Marković (2012: 216), Jelaska (2003) i drugi smatraju da su podjele prema prezentskoj osnovi *ekonomičnije* i logičnije od drugih podjela pa će, oslanjajući se na spomenute radove, testirani model biti temeljen na prezentskoj osnovi. Marković (2012: 226) ističe prednosti Jelaskine podjele, međutim, kako ona i dalje nije dio *jezikoslovnog mainstreama* zahvalnije je evaluirati ustaljene podjele. Odabir je određenog modela, osim iznesenih argumenata, posve arbitraran. Naravno, i dalje moramo imati na umu da je i gramatika Silić – Pranjković (2005) nesavršena, štoviše otvara niz problema, klasifikacijskih dvoumica i nedosljednosti. O tome detaljnije piše Marković (2012: 218-227).

3. Računalna lingvistika i obrada prirodnog jezika

Predmet se ovoga rada može smjestiti na granicu dvaju znanstvenih područja: lingvistike i informatike. Pristupi li se predmetu rada (jeziku) prvenstveno iz lingvističke vizure tj. koristeći tradicionalnu lingvističku teoriju, odnosno promatrajući informatiku i računalnu tehnologiju samo kao sredstvo za proučavanje golemog gradiva, govorimo o računalnoj lingvistici. Približi li se jeziku iz smjera informatike, tj. da se jezik promatra kao i svaki drugi objekt u računalnom okruženju, on je samo skup podataka koji se mora što brže i učinkovitije obraditi, može se govoriti o obradi prirodnog jezika (engl. *natural language processing, NLP*). Dva su pristupa bliska i srodna, ali predmetu ipak pristupaju iz različitih aspekata. Vidi se da se dva područja ne podudaraju. Govoreći općenito, može se reći da je računalna lingvistika konzervativniji, a obrada prirodnoga jezika inovativniji pristup robusnoj i *zatvorenoj* interdisciplinarnoj znanosti koja na neki način uključuje prirodni jezik u svoje područje. Razlike se između dviju znanosti teško definiraju jer su spomenuta područja u stalnom dijalogu te ne postoje ustaljene konvencije i pravila za definiranje granica dviju disciplina (predstavnici se dviju znanstevnih disciplina susreću i diskutiraju na istim znanstvenim konferencijama). Čini se da se razlike ipak mogu izoštriti. Računalna je lingvistika puko korištenje računalne tehnologije za rješavanje pojedinih lingvističkih problema, ona naglasak stavlja na lingvistiku, a računalnom se tehnologijom koristi samo da bi ostvarila ciljeve svog istraživanja tj. kratkoročno, jednokratno, ne gledajući širu sliku i mogućnosti drugih primjena programa, odnosno mogućnost njihova razvijanja. Računalna se lingvistika većinom bavi istraživanjima, ne komercijalizira svoje aplikacije (jer one, u užem smislu, i ne postoje). Obrada se prirodnog jezika više fokusira na *umjetnost koja se krije* u rješavanju inženjerskih, računarskih problema pri obradi prirodnog jezika. Obrada prirodnog jezika u prvom redu služi za analizu velike količine podataka u malo vremena i za unapređivanje ljudsko-strojne komunikacije. Rad s velikom količinom podataka do neke mjere podrazumijeva i skalabilnost te mogućnost proširivanja, prvenstveno, količine podataka, ali i računalnog programa (usp. Eisner 2019; Trenthyer.com 2019). Obrada se prirodnog jezika može opisati kao praktični pristup predmetu, *računarskije* nastrojen pristup jeziku i problemima analize jezika u računalnom okruženju.

Nedvojbeno je da stručnjak koji se bavi računalnom lingvistikom ili obradom prirodnog jezika mora biti obrazovan u obama područjima. Informatičar bez lingvističkog znanja neće obraćati pozornost na korijen, nerazdvojive sintagme, uzroke fonoloških (i fonetskih) promjena i slično, dok lingvist bez znanja u informatici neće moći obraditi veliku količinu

podataka na efikasan način (Tadić 2003: 9-12). Jasno je vidljivo kako jednog nema bez drugog. Razlika postoji, ali inzistiranjem na suradnji, a ne na različitosti tih dvaju područja mogu se postići ogromni napreci na području jezičnih tehnologija (Tsuji 2011: 52; Eisner 2019).

3.1. Jezične tehnologije

Tadić (2003) sve tehnologije koje se temelje na obradi jezika i radu s njim zove jezičnim tehnologijama. Jezične su tehnologije neodvojive od jezične industrije koja u razvoju jezičnih tehnologija vidi veliki potencijal. Razvojem obrade prirodnog jezika raste i razvoj i učinkovitost strojnih prevoditelja, sustava za učenje, umjetne inteligencije, višejezičnih platformi, prepoznavanja govora i pisma i dr. Lista je neiscrpna. Iako živimo u svijetu tehnologije, računala, ekrana, zaslona na dodir i programa za prepoznavanje govora i iako iza svih ovih tehnoloških inovacija i nevjerojatnih postignuća stoje računalni inženjeri, komunikacija se i dalje odvija među ljudima ili „za ljude“, a zbog toga se komunikacija, na kraju, uvijek odvija na prirodnom jeziku (Tadić 2003: 20).

Za ovaj bi rad ipak bilo interesantnije promotriti jedan drugi aspekt jezičnih tehnologija, odnosno njihovu podjelu na jezične resurse, jezične alate i komercijalne proizvode. Tadić kaže da su „jezične resursi računalno pribavljene, pohranjene i podržane zbirke jezičnih podataka, a sastoje se ponajprije od korpusa, a potom i od rječnika“ (*ibid*: 28). Jezični se alati mogu promatrati kao sredstvo za uređivanje, normaliziranje jezičnog resursa. Oni se mogu promatrati na svim jezičnim razinama (fonološkoj, morfološkoj, sintaktičkoj itd.). Za ovaj su rad najvažniji alati koji obrađuju riječi na morfološkoj razini tj. lematizatori te POS i MSD tageri. Prema Tadiću (*ibid*: 41) „komercijalni su proizvodi rezultat svake tehnologije pa i jezičnih tehnologija“. U sljedećoj se rečenici nalazi sintagma koja je nedovoljno razrađena, a čini se da je temelj kategorije *komercijalni proizvodi*: „[proizvod koji] se može nabaviti u tradicionalnim prodavaonicama ili putem WWW-dućana“ (*ibid*). Kao primjere autor navodi rječnike, strojne prevoditelje, provjerenike pravopisa ili stila i dr. (usp. *ibid*: 27-44).

Primijeni li se Tadićeva definicija na ovaj rad, lako se da zaključiti da je resurs ovoga rada korpus. Korpus je „skup jezičnih odsječaka odabranih i skupljenih prema eksplicitnim lingvističkim kriterijima s ciljem da čine jezični uzorak“ (*ibid: 28*). Valja obratiti pozornost na definiciju koju Tadić daje na kraju svoje knjige, u pojmovniku. Tamo ističe kako korpus služi za „analizu ljudskog ponašanja (u domeni jezične uporabe)“ i za „empirijsku provjeru

neke jezične teorije“ (*ibid*: 156). Upravo ove dvije sintagme čine srž ovoga rada i jesu motivacija za izradu ovoga projekta.

3.2. Povijest, namjena i vrste korpusa

Korpusna lingvistika i korpusi u današnje vrijeme podrazumijevaju računalizirane korpuse, ali to nije uvijek bio slučaj jer računala, u današnjem smislu riječi, nisu bila prisutna. Već se i prije osobnih računala javljaju istraživanja koja se retrospektivno mogu promatrati kao preteče korpusne lingvistike i rada s korpusima općenito (Reppen i Simpson-Vlach 2010: 89).

Među ranim istraživanjima jezika pomoću korpusa mora se spomenuti Käding koji 1897. izrađuje frekvenciju slijeda slova u 11 milijuna njemačkih riječi. Liste riječi također se izrađuju na početku 20. stoljeća, a 1952. izrađuje se korpusno utemeljena gramatika engleskog jezika. Ubrzanim razvojem kompjuterske tehnologije krajem 1950-ih povećava se korpusni pristup jezičnoj analizi i lingvistici općenito. Valja spomenuti i Roberta Busa (prema nekima jedan od rodonačelnika digitalne humanistike) koji je izradio prve strojno generirane konkordancije (McEnery i Hardie 2013:2-3).

Na razvoj je korpusne lingvistike u velikoj mjeri utjecao američki jezikoslovac Noam Chomsky. U svojim radovima žestoko kritizira korpusnu lingvistiku. Pri tome se prvenstveno oslanja na de Saussureovo učenje o opreci jezika i govora (*langue* i *parole*) i zacrtava lingvistiku kao znanost koja bi trebala proučavati unutrašnji jezik (mogućnost), a ne vanjski jezik (izvođenje). Smatra da se u umu svakog govornika nalazi apstraktno znanje o govorenom jeziku koje omogućava beskonačno slaganje novih jezičnih izričaja neovisno o tome je li on prije bio ostvaren ili nije. Prema tome lingvisti se moraju baviti sposobnošću stvaranja jezičnog izričaja, a ne pojedinačnih ostvaraja, o čemu zapravo svjedoče korpusi. Uslijed učenja Chomskog dolazi do blagog prijelaza s empirijskog na racionalistički pristup jeziku. Međutim, ističući da se i među egzaktnim znanostima nalaze one koje *samo popisuju* i *sakupljaju građu* (astronomija, geologija), znanstvenici su se usprotivili Chomskom i nastavio se razvoj korpusa i korpusne lingvistike. Danas se teško nailazi na područje lingvistike koje ne koristi korpusni pristup svome predmetu (McEnery i Hardie 2013:4-5).

Korpusi se prema sastavu mogu podijeliti na dvije velike kategorije. Ona kojom je započela korpusna lingvistika, zove se opći (engl. *general*) korpus. Korpusi ovakvog tipa nastoje biti reprezentativni uzorak jednoga jezika, autori u njega svrstavaju tekstove različitih žanrova i diskursa kako bi se dobio korpus koji zrcali jezičnu situaciju jezične zajednice. Opći su

korpusi u pravilu veliki. Korpus koji se i danas često spominje i svojevrstan je prototipni primjer općeg korpusa nekoga jezika jest British National Corpus koji je sastavljen 90-ih i sadrži 100 milijuna riječi/pojavnica (Reppen i Rita Simpson-Vlach 2010: 91). Pri radu se s reprezentativnim općim korpusima uvijek mora imati na umu da su oni uvijek rezultat odabira. Odabirući jezične materijali, sastavljači moraju uzeti u obzir raznolikost žanrova, jezika, autora (njegovog socijalnog, ekonomskog i kulturnog položaja) i dr. U hrvatskoj se lingvistici prvi pokušaj izrade reprezentativnoga korpusa veže za ime Milana Moguša. Projekt *Korpus suvremenog hrvatskog književnog jezika* pokrenut je 1976. s ciljem omogućavanja sustavno proučavanje hrvatskoga jezika. *Mogušev korpus* broji milijun pojavnica i sastavljen je od 5 potkorpusa različitih tekstovnih žanrova (drama, novine, proza, stihovi, udžbenici) (Tadić 1997: 389-390).

Drugu veliku skupinu korpusa čine specijalizirani korpusi. Oni su manjeg opsega i obuhvaćaju samo određene jezične idiome, jezik neke struke, znanstvenog područja i sl. Čini se da suvremena korpusna lingvistika favorizira specijalizirane korpuse jer oni daju cjelovitiju sliku nekoga jezičnog idioma te mogu ukazati na razilaženje značenja pojedinih riječi u različitim kontekstima. Specijalizirani korpusi mogu biti razni. Danas se u velikom broju izrađuju učenički korpusi koji se sastavljaju od tekstova učenika nekoga stranog jezika. Ovakvi korpusi mogu dati uvid u poteškoće pri usvajanju nekog jezika, upozoriti na manjkavosti modela za podučavanje i sl. (Reppen i Simpson-Vlach 2010: 90-92).

Valja posebno istaknuti mrežne korpuse koji su sastavljeni od tekstova preuzetih s mreže. Ovi su korpusi zanimljivi jer ne prolaze ljudsku selekciju, odnosno u njih ulaze svi registri koji se mogu naći na mreži. Ovi korpusi nisu reprezentativni, ali su golemi i stoga su prikladni za analizu jezika, pretraživanje jezičnih konstrukcija i pravila u jeziku. Međutim sastavljači mrežnih korpusa ne mogu jamčiti kvalitetu i reprezentativnost tekstova te njihovu vremensku stabilnost (sastavljači ne znaju hoće li stranica biti dostupna i za par godina). Pri izradi, sastavljači moraju poštivati autorska prava, a to često znači da će pojedini žanrovi i tipovi diskurza, poput književnih djela, biti izostavljeni iz korpusa (Spousta 2006: 180). Osim navedenih tipova korpusa možemo govoriti i o sinkronijskim i dijakronijskim korpusima, o jednojezičnim i paralelnim korpusima, o korpusima pisanog i govornog jezika i dr. (*ibid*: 179).

Svaki korpus, da bi doista bio koristan za rad, mora biti označen. Pri označavanju, pojedinim se oznakama opisuju gramatička svojstva riječi. Što je opis detaljniji, to je veća iskoristivost

korpusa. Nakon označavanja proučavatelji mogu pretraživati korpus i ukazati na osobine nekoga jezika, na neke pravilnosti. Rad s korpusima, bez računalne tehnologije, nezamisliv je. Ipak, bez ljudskog nadzora računalo ne može pružiti zadovoljavajuće rezultate jer ono ne može interpretirati informacije. Korpusna lingvistika može dati uvid u jezičnu materiju jer spaja kvantitativno (računalno) i kvalitativno (ljudsko) (Reppen i Simpson-Vlach 2010: 91).

Pomoću korpusa može se doći do informacija koje se tiču različitih jezičnih razina. Mogu se izraditi jednostavne liste riječi, ali i vrlo složene gramatičke strukture koje omogućavaju interaktivnu analizu lingvističkih i paralingvističkih elemenata. Jedna od najčešćih informacija koje pruža korpus jest frekvencija riječi. Osim frekvencije i konkordancija pojedinih riječi mogu se proučavati i podaci o skupovima riječi, o vjerojatnosti pojavljivanja jedne riječi pored druge i sl. Analizom korpusa može se doći do neznanih činjenica o nekom jeziku te tendencija i modela u njemu[7] (*ibid*: 96).

3.3. Hrvatski mrežni korpus – hrWac

Obrada prirodnog jezika pretpostavlja korpusni pristup jeziku. Drugi je način, posve teorijski, ne-empirijski, za ovo područje neprihvatljiv. Korpusi pružaju uvid u jezičnu građu, a oni s kojima se bavi obrada prirodnog jezika označeni su i pretraživi. To je iznimno važno za ovu znanstvenu disciplinu jer segmentiran, tokeniziran, lematiziran i potpuno označen korpus omogućava njegovo pretraživanje te zadavanje različitih, specifičnih upita (npr. uzimanje u obzir samo glagole ili samo riječi na „n“ i sl.) pri radu s njim. Ovakvi se korpusi lako obrađuju jer su izrazito fleksibilni i brzo odgovaraju na upite. Upiti se lako mogu mijenjati, prilagoditi novim korisničkim ili istraživačkim zahtjevima, te se rezultati, ako je to potrebno, mogu prikazati u usporedbi s drugim rezultatima, a po potrebi i grafički. Korpusi su vrlo prikladni za jezičnu analizu jer je rezultate pretraživanja uvijek moguće vidjeti u kontekstu. Takvo se pretraživanje obično naziva pretraživanjem konkordancija (SketchEngine 2019). Računalno obrađeni korpusi pružaju i druge korisne opcije. Aplikacija Sketch Engine-a omogućava izradu frekvencijskih listi, tezaurusa, izlučivanje ključnih riječi i terminologija, a sve to može učiniti i s paralelnim, višejezičnim korpusima.

Korpus na kojemu se temelji ovaj rad je hrWaC korpus koji je sačinjen od dokumenata prikupljenih s *.hr* domene (*WaC* u nazivu korpusa kratica je za izraz *Web as a Corpus*). U njemu se nalazi 1,4 milijardi pojavnica i oko 90 000 glagola. (Ljubešić i Klubička 2014).

[7] Odabir će sinonima, na primjer, često biti uvjetovan kontekstom.

Korpus hrWaC prvi je korpus ovakvog tipa na hrvatskom jeziku. Paralelno se s njim izradio sličan korpus za slovenski, bosanski i srpski jezik. Označen je, tj. uz svaku se pojavnicu nalaze morfosintaktičke oznake riječi. Uz to, korpus je preko korisničkog sučelja Sketch Engine-a pretraživ. Upit se može postaviti pomoću CQL jezika (*Corpus Query Language*), ali se može postaviti i *jednostavni* upit koji pretražuje lemu ili slijed lema. Ne može se dovoljno istaknuti važnost ovakvog korpusa za razvoj hrvatskih jezičnih tehnologija, korpusne lingvistike, obrade prirodnog jezika pa i pisce preskriptivnih i deskriptivnih jezičnih priručnika. To da su u korpus ušli dokumenti s *.hr* domene i da oni nisu naknadno selektirani znači da se u korpusu uz tekstove pisanim na standardnom hrvatskom jeziku (npr. mrežne stranice Hrvatskog sabora ili drugih službenih i javnih tijela) nalaze tekstovi iz različitih modnih novina, blogova, reklama, korisničkih komentara, forumskih svađa, rasprava i dr. Jezik ovih diskursa bliži je svakodnevnom govoru, a kao takav vjernije zrcali svakodnevno „ljudsko ponašanje“.[8] Ovakav korpus može svjedočiti o živućem žargonu, dvojbi govornika u fleksiji, načinu preuzimanja riječi iz drugog jezika, problemima s (orto)grafijom i *trendovima* u jeziku općenito. Nažalost, ovakav korpus nije reprezentativan. Jedna od najvećih prednosti mrežnih korpusa je njihova veličina i raznolikost tekstova. Iako je s jedne strane tematska raznolikost tekstova prisutna, s druge se strane na diskurzivnoj razini ta raznolikost ne može uočiti. U mrežnim korpusima, prije svega, manjka poetski diskurs. Upravo zbog toga što ne zastupa sve diskurze jednog jezika ravnomjerno, mrežni korpus ne može biti reprezentativan u užem smislu (on i dalje može biti reprezentativan korpus jezika u mrežnom okruženju ili svakodnevnoj pisanoj komunikaciji). Druga je prednost mrežnih korpusa njihova ažurnost, odnosno posve sinkronijski prikaz jezika. Jedna je od najvećih mana mrežnih korpusa nemogućnost dokazivanja autorstva, originalnosti i pouzdanosti pojedinih tekstova (usp. Fletcher 2011).[9] Konkretni se problemi javljaju pri korištenju korpusa. Pretražuje li se korpus, brzo se naiđe na nepravilno napisane riječi. Lematizator (engl. *lemmatiser*), označitelj vrste riječi (engl. *POS tagger*) i morfosintaktički označitelj (engl. *MSD tagger*) ne *raspoznaju* ispravnost neke riječi, ne uspoređuju riječ s nekom ogromnom bazom podataka, ne primjenjuju gramatička pravila. [10,11] Ovi programi obrade i svedu na lemu *svaku* riječ, a zbog toga se u korpusu javljaju nepostojeće riječi i tako netočni podaci ulaze u statistike (pa i one

[8] Ovo se odnosi na definiciju korpusa iz prijašnjeg poglavlja.
[9] Najveći problem predstavlja prepoznavanje spam-a koji često skriven u *lažnim* jezičnim uzorcima.
[10] Prema Tadiću (2003: 32) označitelji vrste riječi programi su koji svakoj pojavnici u tekstu pridjeljuju i podatak o vrsti riječi; morfosintaktički označivač svakoj pojavnici pridjeljuje podatak o vrijednostima ostvarenih morfosintaktičkih kategorija; lematizator svakoj pojavnici u tekstu pridjeljuje njezinu lemu tj. njezin polazni, kanonski, natuknički oblik.
[11] Ovakav bi pristup doduše na implicitan način isto tako *cenzurirao* korpus kao da je on *ručno*, selekcijski sastavljen.

frekvencijske) o korpusu. Sve ovo znači da mrežni korpus uz svoje prednosti ima i svoje mane, stoga svaki zaključak donesen na temelju činjenica iz sličnog korpusa mora biti svjestan svojih manjkavosti i mogućih nedostataka. Na jednom će se primjeru pokazati kako pogrešni primjeri nisu rijetki i ne moraju biti rezultat zatipka (tipfelera).

3.4. Prednosti i mane hrWaC korpusa

Hrvatski govornici, ovisno o regiji u kojoj žive ili u kojoj su rođeni, često koriste riječ *nemrem*. Ona se koristi kada se želi izraziti da nešto govorni subjekt *ne može*.[12] Kako bi postojao oblik *nemrem*, prvo se morao provesti rotacizam u (doduše nestandardnoj) obličnici *možem* tj. dobiti oblik *morem*. Nakon toga niječna se čestica *ne* stopila s glagolom, glas je *o* ispao i dobio se glagol *nemrem*. Glagol *nemrem* u korpusu se pojavljuje 8 525 puta. Njega su programi za prepoznavanje leme i morfosintaktičkih oblika u odredili na različite načine. Primjerice, na stranici jednog autoservisa stoji komentar: „Kaj da radim ????? nemrem bez brisaca“.[13] U ovom je kontekstu prema programima *nemrem* imenica srednjeg roda u instrumentalu jednine. Zbog česte konstrukcije IMENICA+BEZ+IMENICA tu je riječ *nemrem* klasificirana među imenice i dodijeljena mu je lema *nemre*. Češći su primjeri u kojima je riječ *nemrem* određena kao glagol u prvom licu jednine prezenta, a čija je lema *nemrijeti*.[14] Kako će proučavatelj pristupiti ovakvoj pojavi? Situacija se ne može promatrati kao rezultat nemara, korisničke pogreške ili nešto slično, riječ je namjerno ovako napisana. Trebaju li se riječi poput ove uvrstiti u rječnike ili spomenuti u studijama o hrvatskom jeziku? Ovaj rad ne teži dati odgovore na ova pitanja, postavlja ih kako bi ukazao na pitanja i probleme koji se javljaju pri radom s korpusima ovakvog tipa.

Drugu veliku poteškoću predstavljaju istopisnice, odnosno homografi. Istopisnice podrazumijevaju dvije ili više riječi koje imaju isti oblik, a različito značenje. Prema Tadiću razlikujemo dva tipa istopisnosti (*ibid*: 126). Prva je unutrašnja istopisnost koja podrazumijeva da jedna pojavnica može imati više morfosintaktičkih opisa, a da svi opisi

[12] Čini se da je glagol *nemrem* defektivan. Prema Markoviću (2012: 75) „defektivna je ona leksička jedinica kojoj nedostaju pojedini gramatički oblici koje tipično imaju članovi njezine vrste, odnosno nema ostvarenu punu paradigmu koju imaju ostali pripadnici njezine vrste”. Glagolu *nemrem* nedostaje infinitiv, a malo je vjerojatno da će hrvatskim govornicima oblici *nemreš, nemre, nemremo, nemrete, nemru* biti toliko prototipni i samorazumljivi kao oblik za prvo lice jednine. Ipak, u korpusu se čak i za oblik prvog lica množine (*nemremo*) može naći 764 potvrda.

[13]Komentar je 2013. godine bio dostupan na mrežnoj stranici: http://www.autoservis-meic.hr/savjeti?proizvodac=&submit=trazi&start=3100.

[14] Drugi je problem što ovaj glagol ovako ulazi u glagolsku vrstu koja se tradicionalno promatra kao „neplodna“. To su glagoli tipa *umrijeti-umrem.*

budu opisi iste leme.[15] Druga je vrsta istopisnosti vanjska istopisnost koja podrazumijeva da jedna pojavnica može značiti oblike različitih lema. Naravno obje su istopisnosti prisutne i u jeziku i u korpusu. Iako za ovaj rad ovaj problem nije značajan, on se mora spomenuti jer *produži li se korak dalje* s istraživanjem i krene li se proučavati aorist, odmah se nailazi na probleme:

	2. lice jednine	3. lice jednine
AORIST	reče	reče

Problemi se istopisnošću umnožavaju ako se proučavaju imenske riječi, a posebno je zahtjevno proučavati vanjsku istopisnost.

Vidi se da se u radu s korpusom javlja puno značajnih i teško premostivih problema pa se postavlja pitanje ima li uopće smisla raditi na korpusu i donositi zaključke na temelju analize korpusa.[16] Naravno da ima, a tome je više razloga. Prvo, ovakva analiza ne zahtijeva ni veliku financijsku potporu, a s njom se može štedjeti i na ljudskim resursima.[17] Drugo, na ovaj se način dolazi do količine podataka koja je do prije 30 godina bila nezamisliva. Treće, ovakvi korpusi svjedoče o *živom* jeziku. Razgovorni je jezik pun žargona, kratica, anglizama, a u novije se vrijeme i emotikoni ili smajlići mogu i moraju proučavati jer je neupitno da su dio naše svakodnevne komunikacije.[18] Nije upitno da gramatike i pravopisi moraju biti (i) preskriptivni, ali uvid u ovakvu količinu podataka može pridonijeti i lakšem opisu jezika, odnosno deskripciji. Četvrto, metode koje se primjenjuju u ovom radu, mogu se primijeniti i na drugim računalnim korpusima. Izgradi li se opširni korpus hrvatskog standardnog jezika na njemu će se s pomoću ponuđenog modela isto tako moći proučavati glagolske vrste.

Svaki se problem, naravno, da riješiti dodavanjem koda u program, ali pri pisanju ovakvog programa vrlo je važno izgraditi *dobro apstrahiran* sustav. Bude li program prekonkretan, nedovoljno apstraktan, možemo se pitati o korisnosti takvog programa. Možda je takav program lakše zamijeniti s jednim jezikoslovcem koji će ručno obraditi glagole. Međutim, bude li program preapstraktan, objekti proučavanja će mu bježati u krive razrede, jezična će građa izmicati i biti beskorisna. Mora se naći *zlatna sredina*. Računalni program neće moći pratiti nove jezične trendove, neće prepoznati jezične pojave koje su za govornika

[15] Vjerojatno bi bilo *stručnije* govoriti o morfološkoj homonimiji ili sinkretizmu (prema Marković 2012:48-49), ali radi dosljednosti donosi se Tadićeva podjela.
[16] Naravno, misli se na korpuse koji su poput hrWaC-a.
[17] Što je danas, pogotovo na području humanistike, vrlo važan faktor.
[18] Može se spomenuti i najnovija aplikacija pod nazivom *Animoji* koja dopušta generiranje personaliziranog emotikona s pomoću prednje kamere pametnog telefona.

samorazumljive.[19] Ljudski je nadzor nužan i opravdan. Čini se da je ispravno reći i to da bi računalni lingvisti ili informatičari koji se bave obradom prirodnog jezika trebali biti lingvistički obrazovani jer će im inače premetanje, homonimija ili neka druga jezična pojava biti puka slučajnost, *jezični ludizam*, a ne jezikoslovna činjenica.

4. Klasifikacija glagola u glagolske vrste u korpusu hrWaC

U prijašnjim su poglavljima izneseni problemi koji se vežu za korpus, podjelu glagola u glagolske vrste, i prednosti korpusnog pristupa jeziku. U ovom će se dijelu rada opisati postupak razvrstavanja glagola. Ovim se radom želi provjeriti istinitost glagolskih klasifikacija jer izgleda da su hrvatski jezikoslovci donosili zaključke o brojnosti i prototipnosti glagolskih vrsta bez uvida u razmjerno velike korpuse. Ovaj bi rad htio ukazati na važnost numeričkih podataka u lingvistici, na obradu prirodnog jezika kao disciplinu koja je prikladna za takav tip istraživanja te na važnost korištenja numeričkih podataka pri pisanju jezičnih priručnika.

4.1. Definiranje korpus glagola i definiranje frekvencijskih skupina

Korpusu hrWaC može se pristupiti pomoću Sketch Engine-a. Opcija se beskonačnih lista riječi plaća pa se u radu moralo posegnuti za nekom alternativom koja bi mogla simulirati cjelokupni korpus. Uzorak je morao sadržavati i najčešće pojave u jeziku, ali i pojedinosti, rijetkosti u njemu.[20] Reprezentativni se uzorak izradio trojnom podjelom glagola iz cjelokupnog korpusa na najfrekventnije, srednjofrekventne i rijetke glagole. Frekvencijska se lista glagola izrađivala na temelju pojavljivanja glagola u bilo kojem obliku. To znači da se nije koristila frekvencijska lista glagola u prezentu ili nekom drugom glagolskom vremenu ili načinu, nego frekvencijska lista svih mogućih oblika. Ovako se dobila *najobjektivnija* frekvencijska lista. Svaka od triju frekvencijskih kategorija sadrži tisuću različnica. Skupinu najfrekventnijih glagola čini *vrh* frekvencijske liste. Najfrekventniji glagol u korpusu je, kako se i može pretpostaviti, *biti*. Glagol se *biti* u korpusu pojavljuje 3 507 469 puta i njime počinje skupina najfrekventnijih glagola, a završava s glagolom *protezati* koji se javlja 825 puta. Najfrekventniju je skupinu vrlo lako definirati, međutim ostale nije. Ne može se odrediti

[19] Zanimljiv je primjer metateze kao rječogradnog postupka. U slengu će se često pojaviti premetanje koje računalni program ne može prepoznati, npr. *žišku rista* (*kužiš stari*) (usp. Marković 2012: 91).
[20] U aplikaciji SketchEngine-a postoji opcija zadavanja parametra pri pretraživanju. S jednim od tih parametra može se odrediti koji je minimalni broj pojavljivanja leme i tako isključiti pretraživanje lema koje su rezultat zatipaka ili slučajnosti.

srednjofrekventa skupina ako ne znamo *granice* korpusa. Najlogičnije je odrediti rijetku skupinu, a zatim je lakše odrediti skupinu između najfrekventnije i rijetke. Pri određivanju rijetke skupine mora se imati na umu da se u korpusima mogu nalaziti riječi koje se pojavljuju svega jedanput. Ta se pojava u korpusnoj lingvistici naziva *hapax legomenon*. U korpusu hrWaC je 4.296.047 lema (u ovom slučaju možemo govoriti i o riječi) koje se javljaju samo jednom, a 8.153.180 lema koje se javljaju do 5 puta. Kako bi se izbjegla inkorporacija pojedinačnih primjera u statistiku, u ovom radu skupinu rijetkih glagole čine glagoli koji se javljaju od 18 do 25 puta. Razlog tomu da donja granice kategorije nije dva, tri ili pet je činjenica da se u korpusima ovakve veličine isti zatipci i iste greške javljaju više puta. Čak se i u skupini najfrekventnijih glagola nalaze glagoli koji su rezultat korisničkog pojednostavljenja, a koja ne mogu ući u statistiku ovoga rada. Tako će se na primjer glagol *željeti*, zbog toga što govornici hrvatskoga jezika, pišući na mreži ili na svojim pametnim telefonima, često izostavljaju dijakritičke znakove, pojaviti kao *zeljeti* 990 puta. Među najfrekventnijim se glagolima može naići i na glagol *zeliti* koji je problematičan zbog nepisanja dijakritičkih znakova, ali i zbog uporabe netočnog derivacijskog morfa.[21] Granice rijetke skupine definirane su ovako jer se dobila skupina od 1 000 glagola i nije bilo potrebno nasumično izbaciti neke glagole, te se ovakvim određivanjem granica broj *netočnih* glagola smanjio u odnosu na samo dno frekvencijske liste.[22] Definirane su granice najfrekventnije i najrjeđe skupine. Logično je da je između njih, tj. između 825 i 25 pojavljivanja, srednjofrekventna skupina. Odredimo li središnju skupinu prema aritmetičkoj sredini dviju granica (npr. glagoli koji se pojavljuju od 375 do 425 puta) dobit ćemo skupove glagola od svega par stotinjak članova. Kako bi rad metodički bio dosljedan svakako je bilo nužno odabrati skup glagola koji ima približno 1 000 glagola, a koji i dalje vjerno odražava srednjofrekventnu skupinu. U srednju su skupinu ušli glagoli koji se pojavljuju između 375 i 140 puta.[23]

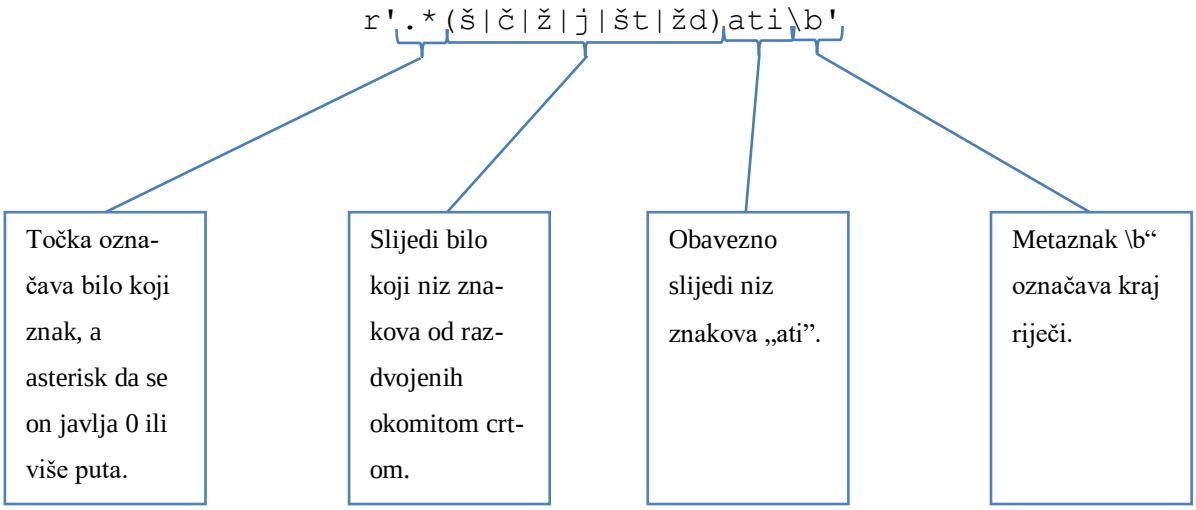
[21] Glagol *željeti* pripada 2. razredu IV. glagolske vrste, tj. tipu *vol-je-ti – vol-i-m* (prema Silić-Pranjković (2005)). Hrvatski govornici analogijom prema prezentskim oblicima (*želim*) tvore infinitiv *želiti, voliti* i slično.

[22] Treba imati na umu da je 1000 limit za korisnika u aplikaciji SketchEngine pa u slučaju da se definiraju granice među kojima je mnogo više od 1000 glagola neki bi se glagoli iz te skupine trebali izbaciti, a to bi metodički bilo neopravdano.

[23] Upitno je je li ovo najbolji izbor za srednjofrekventnu kategoriju, ali nažalost korisnici nemaju uvid u realnu frekvencijsku listu na kojoj bi se onda, znajući točan broj glagola i njihove frekvencije, mogla ustvrditi potpuno objektivna sredina. Vjerojatno bi ona bila puno bliže donjoj granici (od 25 pojavnica).

4.2. Postupak klasifikacije

Nakon definiranja triju skupina slijedila je njihova obrada. Pomoću SketchEngina i opcije *wordlist* dobile su se samo leme i njihova frekvencija. Za određivanje glagolskih vrsta uz infinitiv (koji se podudara s oblikom leme) nužan je i prezentski oblik glagola. Liste su se sjedinile i slijedilo je ručno unošenje prezentskih oblika.[24] Dobivena se lista sastoji od 3 000 lema i odgovarajućeg oblika za prvo lice jednine prezenta. Elementi u parovima odvojeni su tabulatorom („\t“) i svaki je par u novom redu („\n“). Programi ne poznaju granice morfema, ne mogu prepoznati korijene i fonološke promjene. Program koji radi razdiobu glagola u glagolske vrste nije naučen da raspozna promjene koje se događaju s riječima, niti je upoznat s morfologijom, već s pomoću fonoloških značajki pojedinih vrsta sortira glagole.[25] Problemu razdiobe pomoću računalnog jezika moglo bi se pristupiti na više načina, ali se čini da je ponuđeni model najrazumljiviji za početnike jer sadržava jednostavne koncepte.[26] Najjednostavnije rečeno, rad uspoređuje svaki oblik riječi s određenim obrascima. Obrasci su definirani pomoću regularnih izraza (engl. *Regular Expressions, RegEx*).



Na gornjem se primjeru vidi regularni izraz koji odgovara infinitivu trećeg razreda četvrte vrste tj. glagolskom tipu *držati-držim*. Pokušamo li uklopiti infinitiv *držati* u gornji obrazac vidjet ćemo je on ispravan, odnosno istinit.

[24] Ručno se unošenje moglo zamijeniti i s metodom da se lista od glagola poveže s hrLeX sustavom, međutim, to je dugotrajan proces (hrLeX je izrazito velik) i dobar bi se dio glagola prije povezivanja trebao ručno obraditi (ispraviti zatipke, netočno napisane riječi, sortirati neispravne riječi sl.).
[25] Za obradu jezika se koristio programski jezik Python 3.7.2.
[26] Pod drugim se pristupima misli na model koji bi raspoznao korijene i derivacijske afikse i komparirajući ih u infinitivu i prezentu odredio glagolsku vrstu (slično alatu CroDeriv).

Za određivanje je glagolske vrste nužno svaki glagolski par (infinitiv-prezent) usporediti s odgovarajućim obrascima. Tako će npr. obrazac za prezent trećeg razreda četvrte vrste (gore opisani izraz infinitivni je obrazac te kategorije) izgledati ovako: `r'.*(im)\b'`. Program glagole svrstava u određenu vrstu ako infinitivni i prezentski oblici odgovaraju obrascima, znači ako je rezultat obaju obrazaca istinit.

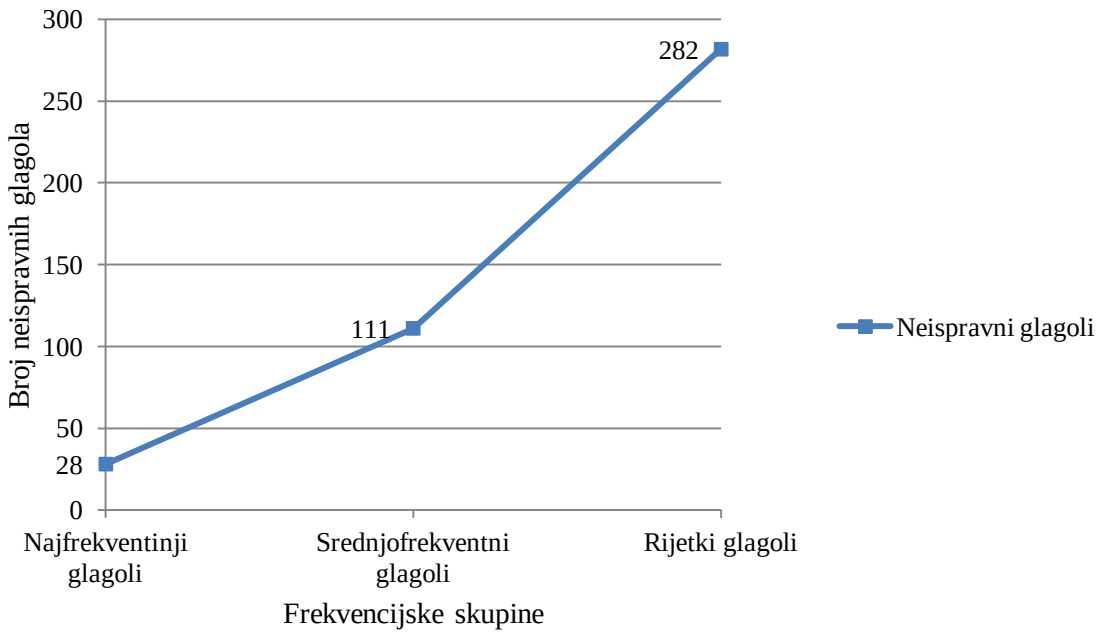
Program uzima svaki glagolski par i uspoređuje ga s obrascima. Ako glagolski par odgovara jednom paru obrazaca, program par svrsta u odgovarajuću vrstu i briše ga s liste na kojoj su svi glagoli. U programu se prije provjerava pripada li glagol drugom razredu pete vrste ili prvom razredu pete vrste. Svi glagoli drugog razreda pripadaju i prvom razredu (po obliku). Drugi razred čine glagoli tipa *proučavati-proučavam*, a prvi razred glagoli tipa *kopati-kopam*. Vidimo da bi glagol *proučavati* odgovarao i obrascu prvog razreda pete vrste koji definira samo obavezno „ati" na kraju riječi. Bilo je nužno držati se spomenutog redoslijeda i nakon toga brisati glagolske parove koji su zadovoljili kriterijima jer bi se u suprotnom glagoli tipa *proučavati* mogli naći u netočnoj grupi ili u više grupa.

Na kraju analize i razvrstavanja program otvara novu tekstualnu datoteku i ispisuje glagolske parove, brojčanost glagolskih razreda i glagole koje nije uspio smjestiti. Među nesvrstanim se glagolima nalaze *nepravilni* glagoli, te ta kategorija može služiti kao indikator manjkavosti nekog modela.[27]

4.3. Neispravni glagoli

Neispravni su glagoli, u ovome radu, oni koji nisu prave hrvatske riječi, odnosno netočno su napisane riječi. Prije negoli je došlo do svrstavanja, glagoli su se morali *ručno* provjeriti kako bi bilo sigurno da su kategorizirani glagoli jedinstveni, pravilno napisani, prave su hrvatske riječi. Oduzimanjem nepravilnih glagola dobilo se 2 587 glagola. U skupu od 3 000 glagola bilo je 413 nepravilnih. Uklanjanje je bilo nužno jer bi se inače dobili neispravni statistički podaci. U svim je trima frekvencijskim skupinama bilo neispravnih glagola, ali kao što se i može očekivati, micanjem od frekventne skupine prema manje frekventnoj, broj se nepravilnih glagola povećavao. Na sljedećem se grafikonu može vidjeti kako se broj neispravnih glagola povećava udaljavanjem od vrha frekvencijske liste.

[27] Vrlo logično, tu će biti glagoli koji ne idu ni u jedan model. Problemi se mogu javiti i drugdje i nije ova jedina kategorija koja može ukazati na probleme s modelom.



Grafikon 1. Broj neispravnih glagola po frekvencijskim skupinama.

Neispravni glagoli čine posebnu skupinu i postavljaju niz pitanja. U ovom se radu proučava živi jezik, ali se ipak odlučilo da se izbace dijalektalizmi koji su isto tako konstitutivni elementi živog, forumskog, svakodnevnog jezika. Takav se postupak opravdava isključivo time da se želi ponuditi komparacija dosadašnjih razmatranja o glagolskim podjelama s dobivenim podacima u ovom radu, te da su sve klasifikacije u glagolske vrste rađene na korpusu standardnoga hrvatskoga jezika.[28] Bilo bi izrazito zanimljivo da se u istraživanje uključe i dijalektalizmi, ali bi se za to trebao izgraditi novi metodički model i dobiveni se statistički podaci ne bi mogli usporediti s dosadašnjim pregledima koji proučavaju standardni hrvatski jezik. Iako se o ovoj problematici može jako puno pisati, čini se da bi detaljnom analizom neispravnih glagola ovaj rad svratio fokus sa svoga predmeta pa će se donijeti samo neki reprezentativni primjeri koji svakako mogu dočarati probleme koji se javljaju pri radu s korpusom.

4.3.1. Kratka tipologija neispravnih glagola

Među najfrekventnijim glagolima zanimljivi su neispravni glagoli koji se vežu za refleks staroslavenskog jata (ě). U standardnoj inačici hrvatskoga jezika jat je u dugom slogu dao dugo „je“ koje se najčešće bilježi slijedom „ije“, a u kratkom je dao kratko „je“ koje se bilježi slijedom „je“. Tako će prema standardu biti *voljeti* (a ne *volijeti*), *lijepiti* (a ne *ljepiti*) i slično. Međutim, među najfrekventnijim se glagolima našao glagol *tribati* (trebati). Posebnu

[28] Ako ostavimo oblike *verujem, virujem* i *vjerujem* u korpusu proučavanih glagola onda se cijela struktura rada mijenja i dovodi se u pitanje objektivnost frekvencijskih podataka.

kategoriju čine glagoli koji su načinjeni od niječne čestice *ne* i glagola. Oni se u neformalnom pismu često pišu neodvojeno pa će se tako među najfrekventnijim glagolima naći *neznati* i *nemoći*. Treći je reprezentativan primjer glagol *neciti* koji se dobio od pomoćnoga glagola *neće*. Izostavio se dijakritik, a lematizator onda nije povezao pomoćni glagol s infinitivom koji ga slijedi i s kojim čini analitički oblik budućeg vremena, te stvorio novu lemu.

Među srednjofrekventim i rijetkim glagolima bezbroj je različitih grešaka. Tipologija tih pogrešaka bila bi preiscrpna pa će se navesti samo nekoliko uzroka netočnog lematiziranja:

- Izostavljanje dijakritičkih znakova (*rijesiti* [riješiti]).
- Netočno lematiziranje riječi (među glagolima se našla imenica *odabir,* a upitno-odnosna zamjenica *ča* se lematizirala kao glagol i generirana je lema *čati*).
- Problemi vezani za kodiranje teksta (*èuvati* [čuvati], *uređivati* [uređivati]).
- Neprovođenje glasovnih promjena (*rasčistiti* [raščistiti]).
- Neprepoznavanje vlastitih imena i lematiziranje istih prema glagolskom obrascu (*agrati* [prema njemačkom nazivu Zagreba Agram]).

5. Podaci o glagolskim vrstama

Marković navodi da se s numeričkim podacima o frekventnosti pojedinih glagolskih vrsta može susresti u svega par gramatika. Navodi gramatiku Babić *et al.* (1991) i Raguževu gramatiku (1997). Uz to spominje i radove Zrinke Jelaske (osobito onaj iz 2003). U ovom će se radu donijeti kratki pregled numeričkih podataka iz tih priručnika.

U Babić *et al.* (1991) mogu se naći sljedeći podaci:

- I. je vrsta ograničena (prema Silićevoj podjeli I. vrsta), vrlo je nevjerojatno da će se u njoj pojaviti novi glagoli.
- Skupina glagola tipa *brinuti-brinem* (prema Silićevoj podjeli II. vrsta) vrlo je brojna i plodna.
- Glagola tipa *vidjeti-vidim* (prema Silićevoj podjeli IV. vrsta 2. razred) svega je šezdesetak s oko 200 izvedenica.
- Glagola tipa *misliti-mislim* (prema Silićevoj podjeli IV. vrsta 1. razred) veoma je mnogo.
- Glagola tipa *stršati-stršim* (prema Silićevoj podjeli III. vrsta 3. razred) je sedamdesetak.

- Skupina glagola tipa *bljuvati-bljujem* (prema Silićevoj podjeli III. vrsta 2. razred) ograničena je.
- Glagola tipa *gledati-gledam* (prema Silićevoj podjeli V. vrsta 1. razred) ima veoma mnogo.
- Glagola tipa *pisati-pišem* (prema Silićevoj podjeli III. vrsta 1. razred) ima mnogo.
- Glagola tipa *putovati-putujem* i *kraljevati-kraljujem* (prema Silićevoj podjeli VI. vrsta 1. razred) ima u velikom broju.
- Glagola tipa *bičevati-bičujem* (prema Silićevoj podjeli VI. vrsta 1. razred) ima svega dvadesetak.

U Raguževoj gramatici (1997) navode se sljedeći podaci:

- Glagola tipa *vidjeti-vidim* (prema Silićevoj podjeli IV. vrsta 2. razred) ima nekoliko stotina.
- Glagola tipa *misliti-mislim* (prema Silićevoj podjeli IV. vrsta 1. razred) ima otprilike 6 000.
- Glagola tipa *budalisati-budališem* (kod Silića se ne uklapa ni u jednu vrstu) ima petnaestak i čine skupinu glagola koja je stilski obilježena i nestandardna.

Prije davanja statističkog prikaza Zrinke Jelaske valja detaljnije proučiti iznesene podatke. Vidi se da su statistički podaci u Raguževoj gramatici skoro posve neprisutni. U njoj se javlja podatak o dvjema vrstama i o jednoj vrsti koja ne opisuje glagole standardnoga hrvatskoga jezika. Ne može se očekivati točan broj pojedinih glagolskih vrsta, ali zbog nedostatka numeričkih podataka drugih glagolskih vrsti izneseni podaci ne djeluju uvjerljivo. Postavlja se pitanje ima li mjesta u „opisu suvremenog hrvatskoga standardnoga jezika“ za opis nestandardnoga. Svakako bi bilo poželjno posvetiti poseban opis nestandardnim izrazima, konjugacijama i sličnom, međutim, u tom bi se slučaju trebali navesti i drugi primjeri koji odudaraju od standarda, a dio su hrvatskoga jezika.

U gramatici Babić *et al.* (1991) vidimo više podataka, ali su oni i u ovom slučaju posve okvirni. Na više se mjesta mogu vidjeti količine poput *mnogo, puno, veoma mnogo*. Koje se brojke skrivaju iza tih količinskih priloga? Na drugim se mjestima spominje da određena skupina glagola ima *dvadesetak, šezdesetak* ili *sedamdesetak* glagola. U obama je slučajevima nužno obratiti pozornost na to da se nigdje ne definira korpus glagola na kojem se izradio

popis.[29] Smatra se da je to problematično za tumačenje pojedinih podataka. Izradimo li korpus dječjeg govora i analiziramo glagolske vrste, u njemu ćemo najvjerojatnije dobiti drugačiju distribuciju nego u govoru odraslih.

Zrinka Jelaska problematizirajući učenje glagola kod govornika kojima je hrvatski drugi jezik donosi 16 000 najfrekventnijih glagola (Jelaska: 2003: 55-57). Ovakva je količina podataka dotada neviđena u hrvatskoj lingvistici i izrazito je korisna. Problem koji se javlja kod ove podjele je svakako korpus na kojem je nastao. Čestotni rječnik hrvatskoga jezika (1999) izgrađen je kao odraz hrvatskoga književnoga jezika, međutim, u njega su uključeni tekstovi samo od 1975. To bi se moglo promatrati i kao proučavanje suvremenog, živog (onodobno) jezika, ali korpus čine sljedeći potkorpusi: drama, novine, proza, stihovi, udžbenici pa se ta mogućnost može odbaciti (Moguš, Bratanić i Tadić 1999: 6). Vidi se da je popis glagola iz ove jezične domene i iz domene internetskog korpusa posve različit. Ne poriče se korisnost ovakvog pregleda, štoviše, revolucionaran je i koristan za komparaciju. Donose se statistički podaci iz tog rada. Broj 100 označava 100 najfrekventnijih glagola, a 16 000 prvih 16 000 najfrekventnijih glagola (Jelaska 2003: 56):

[29] U uvodu *Tvorbe* piše: „...sva tri svoja organska narječja upotrebljavali su Hrvati i kao književni izraz. Na njima su sastavljali bilješke, zapise, književna djela, isprave, pisma i druge tekstove. (...) Prikaz toga prožimanja (*triju dijalekata*) osnovica je ovom povijesnom pregledu“. Nakon ove rečenice slijedi kratki povijesni pregled hrvatskog jezika. Bitno je da se nigdje ne spominje koji su to izvori iz kojih se crpi. Može se pretpostaviti da temelj čini knjiški, novinarski i *razgovorni* jezik.

Vrsta	Razred	100	100 (po vrsti)	16 000 (po vrsti)
a	gledati-gledam	22%	22%	36%
i	moliti-molim	26%	37%	30%
	voljeti-volim	6%		
	bježati-bježim	5%		
e	dignuti-dignem	0%	12%	29%
	vjerovati-vjerujem	4%		
	davati-dajem	1%		
	smijati se-smijem se	2%		
	plesati-plešem	5%		
ø	naći-nađem		29%	5%

Tablica 1. Frekvencija glagolskih kategorija iz Jelaska (2003).

Značajan je i numerički pregled Zrinke Jelaske i Tomislave Bošnjak Botica. Proučavajući prototipnost konjugacijskih kategorije autorice su napravile iscrpnu statistiku o frekvenciji 24 538 glagola (Jelaska i Bošnjak Botica 2019). Statistički podaci istraživanja korisni su i oni će se usporediti sa statističkim podacima koji su dobiveni u ovom radu. U radu nema podataka o korpusu na kojem se izradio popis pa se ne mogu donijeti zaključci o tome koja se jezična domena opisuje. Donose se frekvencijski podaci pojedinih glagolskih kategorija (Jelaska i Bošnjak Botica 2019):

	Reprezentativni glagoli	Frekvencija razreda	Glagolska vrsta	Frekvencija glagolskih vrsta
1	gledati	9590	a	9590
2	moliti	7011	i	7745
3	vidjeti	509		
4	trčati	225		
5	pisati	1325	e1	5813
6	smijati se	337		
7	putovati	2621		
8	davati	67		
9	viknuti	1463		
10	naći-nađem	1390		1390
			e1+e2	7203
Total		24.538		24.538

Tablica 2. Frekvencija glagolskih kategorija u Jelaska i Bošnjak Botica (2019).

	Reprezentativni glagoli	% glagolskih razreda u odnosu na ukupan broj glagola	Glagolska vrsta	% glagolskih vrsta u odnosu na ukupan broj glagola
1	gledati	39.08	a	39.08
2	moliti	28.57	i	32.02
3	vidjeti	2.07		
4	trčati	1.37		
5	plesati	5.4	e1	23.23
6	smijati se	0.92		
7	vjerovati	10.68		
8	davati	0.27		
9	krenuti	5.96		
10	ići	5.66		5.66
			e1+e2	28.9
Total		100		100

Tablica 3. Frekvencija glagolskih kategorija u postocima u Jelaska i Bošnjak Botica (2019).

5.1. Frekvencijska distribucija glagola na glagolskom korpusu iz hrWaC-a

U ovom će se poglavlju donijeti frekvencijski pregled koji je izrađen na već spomenutom korpusu s već opisanim postupkom. Radi lakšeg pregleda izrađene su četiri tablice. U prvoj se tablici mogu vidjeti podaci svih triju frekvencijskih skupina, a nakon toga slijede frekvencijske tablice pojedinih frekvencijskih skupina. U tablicama se uz reprezentativne primjere donose i vrsta i razred glagola (prema gramatici Silić i Pranjković (2005)):

		Reprezentativni primjer	Broj	Broj po vrstama	Postotak po razredima	Postotak po vrstama
I. vrsta[30]		ići	278	278	10,7%	10,7%
II. vrsta		viknuti	96	96	3,7%	3,7%
III. vrsta	1. razred	pisati	143	172	5%	6%
	2. razred	kljuvati	1			
	3.razred	grijati	28		1%	
IV. vrsta	1. razred	raditi	821	897	31,7%	34,5%
	2. razred	vidjeti	49		1,8%	
	3. razred	trčati	27		1%	
V. vrsta	1. razred	kopati	810	952	31,3%	36,2%
	2. razred	proučavati	142		4,9%	
VI. vrsta	1. razred	kupovati	49	187	1,8%	6,6%
	2. razred	smanjivati	138		4,8%	
Sve			2582+5	2582	100%	100%

Tablica 4. Frekvencija glagolskih kategorija u provedenom istraživanju.

	Jelaska (2003)	Jelaska i Bošnjak Botica (2019)	Statistika ovoga rada	
I. vrsta	5%	5,66%	10,7%	
IV. vrsta	30%	32,02%	34,5	
V. vrsta	36%	39,08%	36,2%	
II. vrsta	29%	23,23%	3,7%	16,3%
III. vrsta			6%	
VI. vrsta			6,6%	

Tablica 5. Usporedba triju statistika

[30] Prva se glagolska vrsta nije rastavljala na podrazrede. U gramatici Silić i Pranjković (2005) ona ima 18 razreda. Čini se da bi bilo zalihosno raditi statistiku u kojoj se ti razredi posebno opisuju. Oni su vrlo specifični i često je pripadnost pojedinih glagola nekom razredu upitna. Osim toga za komparaciju će nam biti važni podaci vezani za prvi razred u cjelosti jer druge podjele ne razdvajaju prvu vrstu na toliko razreda.

		Reprezentativni primjer	Broj	Broj po vrstama	Postotak	Postotak po vrstama
I. vrsta		ići	136	136	14%	14%
II. vrsta		viknuti	23	23	2,3%	2,3%
III. vrsta	1. razred	pisati	54	68	5,5%	6,9%
	2. razred	kljuvati	0			
	3.razred	grijati	14		1,4%	
IV. vrsta	1. razred	raditi	358	392	36,9%	40,3%
	2. razred	vidjeti	19		1,9%	
	3. razred	trčati	15		1,5%	
V. vrsta	1. razred	kopati	234	280	24,1%	28,8%
	2. razred	proučavati	46		4,7%	
VI. vrsta	1. razred	kupovati	24	70	2,4%	7,1%
	2. razred	smanjivati	46		4,7%	
Sve			969-3 ostalo	969	100%	100%

Tablica 6. Frekvencija najfrekventnijih glagola po glagolskim kategorijama u provedenom istraživanju.

		Reprezentativni primjer	Broj	Broj po vrstama	Postotak	Postotak po vrstama
I. vrsta		ići	101	101	11,2%	11,2%
II. vrsta		viknuti	34	34	3,7%	3,7%
III. vrsta	1. razred	pisati	49	59	5,4%	6,5%
	2. razred	kljuvati	0			
	3.razred	grijati	10		1,1%	
IV. vrsta	1. razred	raditi	261	284	29%	31,5%
	2. razred	vidjeti	18		2%	
	3. razred	trčati	5		0,5%	
V. vrsta	1. razred	kopati	289	336	32,2%	37,4%
	2. razred	proučavati	47		5,2%	
VI. vrsta	1. razred	kupovati	17	83	1,8%	9,1%
	2. razred	smanjivati	66		7,3%	
Sve			897+2	897	100%	100%

Tablica 7. Frekvencija srednjofrekventnih glagola po glagolskim kategorijama u provedenom istraživanju.

		Reprezentativni primjer	Broj	Broj po vrstama	Postotak	Postotak po vrstama
I. vrsta		ići	41	41	5,7%	5,7%
II. vrsta		viknuti	39	39	5,4%	5,4%
III. vrsta	1. razred	pisati	40	45	5,5%	6%
	2. razred	kljuvati	1			
	3.razred	grijati	4		0,5%	
IV. vrsta	1. razred	raditi	203	222	28,3%	30,8%
	2. razred	vidjeti	12		1,6%	
	3. razred	trčati	7		0,9%	
V. vrsta	1. razred	kopati	287	336	40%	46,8%
	2. razred	proučavati	49		6,8%	
VI. vrsta	1. razred	kupovati	8	34	1,1%	4.7%
	2. razred	smanjivati	26		3,6%	
Sve			717+1	717	100%	100%

Tablica 8. Frekvencija nefrekventnih glagola po glagolskim kategorijama u provedenom istraživanju.

6. Usporedba rezultata

U prvom dijelu ovog poglavlja usporedit će se statistički podaci iz radova Zrinke Jelaske i Tomislave Bošnjak Botice s podacima iz korpusa hrWaC, a u drugom dijelu analizirat će se zanimljivosti među trima frekvencijskim kategorijama iz korpusa hrWaC.

Prvo treba obratiti pozornost na činjenicu da radovi Jelaske i Bošnjak Botice obuhvaćaju više glagola od ovoga rada. Zanimljivo je usporediti ih sa statistikom iz ovoga rada jer, iako on obuhvaća znatno manje glagola, statistika ovoga rada u sebi objedinjuje i vrlo frekventne i nefrekventne glagole i pruža jedinstven uvid u glagolsku distribuciju različitih frekvencijskih skupina. Statistički podaci triju radova ne razlikuju se u velikoj mjeri. To govori o tome da se glagoli iz korpusa hrWaC, najvjerojatnije, u dobrom dijelu podudaraju s glagolima iz Čestotnika, a to ističe važnost proučavanja glagola koji su na dnu frekvencijske liste mrežnog korpusa. Neke se sitne razlike ipak mogu uočiti. Glagoli tipa *gledati-gledam* u svim su podjelama najčešći.[31] Glagolska vrsta s prezentskim nastavcima *i* (*moliti-molim*, ali i *vidjeti-vidim*) druga je najfrekventnija kategorija, ali raspodjela u razrede odudara. Dok u Jelaska (2003) glagoli tipa *voljeti-volim* čine 6% svih glagola, a glagoli tipa *trčati-trčim* 5%, u Jelaska i Bošnjak Botica (2019) i ovom radu glagoli tipa *voljeti-volim* čine otprilike 2% svih glagola, a *trčati-trčim* otprilike 1%. Zanimljivo je i da se u radovima Jelaska (2003) i Jelaska i Bošnjak

[31] U ovom se dijelu rada neće koristiti klasifikacija glagola iz gramatike Silić i Pranjković (2005) jer Jelaska koristi svoju glagolsku klasifikaciju. Smatra se da je puno preglednije i jednostavnije ako se glagolske vrste u ovom dijelu imenuju prema reprezentativnim primjerima pojedinih kategorija.

Botica (2019) vrsta s nultim morfemom (glagoli tipa *gristi-grizem, bosti-bodem* i sl.) čini svega otprilike 5% svih glagola, a u statistici ovoga rada taj se postotak penje čak do 10,7%. Posebna se pozornost mora obratiti glagolskoj vrsti tipa *dignuti-dignem*. Za nju se govori da je vrlo plodna i mnogobrojna (usp. Babić *et. al*. (1991) i Raguž (1997). Tih je glagola u ovom radu i u radu Jelaske i Bošnjak Botice (2019) svega 3,7% i 5,96% u odnosu na sve glagole, dok Jelaska (2003) ne donosi zastupljenost razreda, samo vrste.

Uočeno je da nema velikih razlika u cjelokupnim statističkim podacima. Iako Jelaska (2003) donosi usporedbu između skupine najfrekventnijih glagola i svih analiziranih glagola, čini se da se dosad, osim Jelaske i Bošnjak Botice (2019), pojedine frekvencijske skupine nisu uspoređivale. Zbog toga najzanimljiviji dio statistike iznesene u ovom radu može biti komparacija triju frekvencijskih skupina. Vidi se da je u svim statistikama glagolska vrsta *misliti-mislim* među najfrekventnijim glagolskim vrstama. U najfrekventnijoj kategoriji ta je glagolska vrsta najčešća s 40,3%, a slijedi je skupina glagola tipa *gledati-gledam* sa svega 28,8%. Krećući se od frekventnijih glagola prema manje frekventnim, omjer se dviju kategorija mijenja. U skupini rijetkih glagola, glagoli tipa *misliti-mislim* čine samo 30,8%, a glagoli tipa *gledati-gledam* već 46,8%. Riječ je o dosta velikom pomaku koji svakako govori o prototipnosti i plodnosti vrste *gledati-gledam* jer su se među najmanje frekventnim glagolima našli glagoli poput *odblokirati* [ukloniti blokiranje s prijatelja na društvenim mrežama?], *štrumpfetati* [ponašati se kao Štrumpfeta iz crtića Štrumpfovi?; dati se olako u spolni odnos?, usp. *štrumfeta*], *rentati* [iznajmiti]. Svi su ti glagoli posve logično ušli u glagolsku vrstu tipa *gledati-gledam*. Puno je vjerojatnije da će biti *štrumpfetati*, nego *štrumpfetjeti*. Očekivano je i to da će postotak glagola s nultim morfemom padati s rastom broja glagola (usp. Jelaska (2003), Jelaska i Bošnjak Botica (2019)), a činjenica da je i među nefrekventnim glagolima značajan broj tih glagola (čak 5,7%) ukazuje na to da nepredvidivost nije povezana s frekventnošću.[32] Među glagolima s nultim morfemom u nefrekventnoj su se kategoriji našli glagoli poput *rastresti, prigristi, štići* (!), *crpsti* itd. Pretpostavljena frekventna glagolska vrsta *viknuti-viknem* dosta je rijetka u svim trima frekvencijskim skupinama i čini svega 2,3%, 3,7% i 5,4% glagola. Valja obratiti pozornost i na 1. razred VI. vrste (*kupovati-kupujem*). Prema Babić *et al* (1991) tih glagola *ima u velikom broju,* ali frekvencijska analiza ovoga rada pokazuje da ih je svega 49 u proučavanom korpusu, odnosno čine 1,8% glagola.

[32] Ovo se ističe jer bi se zbog velike zastupljenosti nepravilnih glagola (u širem smislu, tj. uključujući glagole s nultim morfemom) na samom vrhu frekvencijske liste moglo pomisliti da su svi netipični glagoli frekventni.

Zbog vidnih parnjaka tih je glagola doista puno, međutim, u živom se jeziku, izgleda, malo koriste.

U statističkim se podacima može vidjeti da se analiziralo 2582+5 glagola. Dodatnih pet glagola čine glagolsku vrstu *nepravilnih glagola*. Oni se ne mogu svrstati ni u jednu od glagolskih vrsta. Nepravilni su glagoli *biti-jesam, moći-mogu, spati-spim, zaspati-zaspim* i *htjeti-hoću*.

7. Zaključak

U ovom su se radu nakon uvoda o morfologiji glagola, glagolskim vrstama, korpusu i metodologiji rada iznijeli novi podaci o frekvencijama pojedinih glagolskih vrsta. Vidjelo se da postoje brojnije i sveobuhvatnije statistike, međutim, ovaj je rad jedinstven jer je ponuđena statistika izgrađena na mrežnom korpusu i zbog toga zrcali živi jezik, jezik s foruma, *jezik naroda*. Opažanja pri komparaciji dobivenih rezultata s rezultatima prijašnjih statistika pokazala su neka odudaranja čiji razlog u ovom radu nije istražen, a svakako bi bilo vrijedno istražiti uzroke razilaženja podataka. Ipak, važno je istaknuti kako jezični uzorak od svega 3 000 glagola ne može zrcaliti pravu sliku jezika, čak ni onda ako se definiraju frekvencijske skupine. Rezultati se stoga moraju razmatrati sa zadrškom. Čini se da bi ovakav pristup proučavanju glagolskih vrsti mogao pružiti nove uvide jezikoslovcima i da bi statistika izrađena na većem jezičnom uzorku ukazala na zanimljive podatke vezano uz glagolsku klasifikaciju. Iako je cilj ovoga rada bio reevaluacija prijašnjih numeričkih podataka o frekventnosti pojedinih glagolskih vrsta, on prije svega ukazuje na važnost obrade prirodnog jezika pri proučavanju lingvistike. Ovakvim se pristupom u vrlo malo vremena može obraditi ogromna količina podataka, a numerički podaci danas, u sumraku humanistike, blago su svim lingvistima.[33] Snaga je korištenog modela ili programa, a i drugih alata za obradu prirodnog jezika, u tome što su oni ponovno upotrebljivi. Izradi li se značajan, veliki korpus suvremenoga hrvatskoga jezika, ovom će se metodom u vrlo kratkom periodu moći donijeti zaključci o distribuciji glagola u glagolske vrste.

U radu su se opisali i nedostaci korpusa, odnosno problemi koji se javljaju pri radu s njim. Često se u suvremenoj lingvistici iznose samo pozitivne strane korpusne lingvistike, njezine neupitne prednosti. One se nikako ne mogu poreći, ali bi bilo nužno obratiti pozornost na nedostatke tih jezičnih resursa. Tu se naravno mora imati na umu stanje hrvatskih korpusa u odnosu na engleskih ili korpusa drugih velikih jezika. Posve je jasno da korpusi izrađeni za svega par milijuna govornika nisu dovoljno ažurni i da se nakon inicijalne izrade malo vremena i truda ulaže u njihov daljnji razvitak. Želi li se hrvatski jezik držati trendova lingvistike poboljšanje bi sustava bilo nužno.

Među nedostacima su se istaknuli glagoli spomenuti pod poglavljem *Neispravni glagoli*. Toj bi se temi mogao posvetiti jedan cijeli rad. Ova skupina uz jezične navike govornika hrvatskoga jezika govori i o problemima lematizatora, o sintagmama u kojima se programi ne

[33] Iako se lingivstika nikako ne smije nasilno predstavljati kao egzaktna znanost jer ona to nije.

snalaze, ukazuje na činjenicu da lematizator ne razumije jezik i da se, koristi li se u dinamičnom okruženju poput interneta, mora s vremena na vrijeme ažurirati i prilagoditi novom kontekstu. U hrvatskome su pisanom jeziku mreže često javljaju riječi koje se pišu bez dijakritičkih znakova i one predstavljaju vrlo veliki problem u ovakvim, posve inkluzivnim, korpusima. Pitanje je kako se taj problem može riješiti, a da cijeli proces ostaje automatiziran. Primjeri su među *neispravnim glagolima* ukazali na niz nedostataka pri računalnoj obradi teksta i na činjenicu da svaki skup podataka koji je preuzet iz sustava i koji je programski obrađen mora proći i ljudsku analizu. Ono na čemu se može poraditi je pojednostavljivanje analitičarevog zadatka.

Iz statistike je izostavljeno 413 glagola, a njihovo je izostavljanje opravdano njihovom neispravnošću. Je li ovakva cenzura korpusa slična prijašnjim opisima hrvatskoga jezika? Uspiju li se jednom ažurirati lematizatori, parseri, morfosintaktički označitelji i drugi programi za obrađivanje prirodnog jezika i postignu li ti alati tada *prihvatljive* rezultate, izbacivanje riječi pod krinkom *neispravnosti* postat će vrlo upitan postupak pri analizi ovakvog tipa korpusa. Hoće li se riječi bez dijakritika smatrati posebnim riječima ili će oznaku *neispravni* izbjeći samo glagol poput *nemrijeti*?

Promatramo li rad u njegovu širem kontekstu, on ističe kako je računalna obrada jezika metoda koja je neupitni dio lingvistike, ali i da nije savršena, da programi ne razmišljaju i da se do provjerenih statističkih podataka dolazi ljudskom analizom. Budućnost se lingvistike, izgleda, temelji na interdisciplinarnom pristupu jezičnom resursu pa stoga struka mora prihvatiti izazove i inkorporirati računarstvo u svoje područje interesa.

Literatura:

Babić, S., Pavešić S & Težak, S. (1991). Oblici hrvatskoga kknjiževnog jezika. U Stjepan Babić *et al*. (1991) *Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika: Nacrti za gramatiku*. (pp 453-741). Zagreb: HAZU-Globus.

Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V., & Znika, M. (1995). *Hrvatska gramatika*. Zagreb: Školska knjiga.

Bošnjak Botica, T. (2013). Opća načela podjela na glagolske vrste u hrvatskome u perspektivi drugih bliskih jezika. *Lahor*, 1 (15), 63-90.

Croatian Web (hrWaC 2.2, RFTagger). Korišteno na https://www.sketchengine.eu/ (05.05.2019)

Eisner, J. (2019). *How is computational linguistics different from natural language processing?*. [mrežna stranica] Quora. Dostupno na: https://www.quora.com/How-is-computational-linguistics-different-from-natural-language-processing (17.06.2019)

Fletcher, W. H. (2012). Corpus analysis of the world wide web. *The encyclopedia of applied linguistics*.

Jelaska, Z. (2003). Proizvodnja glagolskih oblika hrvatskoga jezika kao stranoga jezika: od infinitiva prema prezentu. U Botica S. (ur.), *Zagrebačka slavistička škola 2002.* (pp 48-63). Zagreb: FFpress Filozofski fakultet

Jelaska, Z. & Bošnjak Botica, T. (2019). Conjugational Types in Croatian. *Časopis instituta za hrvatski jezik i jezikoslovlje*, 45 (1), 47-74.

Ljubešić, N. & Klubička, F. (2014). {bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian. *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*. (pp. 29-35). Gothenburg: Association for Computational Linguistics. Preuzeto s https://www.aclweb.org/anthology/W14-0405 (05.05.2019)

Ljubešić, N. & Klubička, F. (2016). *Inflectional lexicon hrLex 1.0,* Slovenian language resource repository CLARIN.SI. Preuzeto s: http://hdl.handle.net/11356/1056 (05.05.2019)

Marković, I. (2012). *Uvod u jezičnu morfologiju*. Zagreb: Disput.

McEnery, T., & Hardie, A. (2013). The history of corpus linguistics. U Keith A. (ur.), *The Oxford handbook of the history of linguistics* (pp. 727-745). Croydon: Oxford University Press.

Moguš, M., Bratanić, M., & Tadić, M. (1999). *Hrvatski čestotni rječnik*. Zagreb: Školska kniga - Zavod za lingvistiku Filozofskog fakulteta.

Python Software Foundation. *Python Language Reference, version 3.7.2.*. Preuzeto s: http://www.python.org (10.04.2019)

Raguž, D. (1997) *Praktična hrvatska gramatika*. Zagreb: Medicinska naklada.

Reppen, R. & Simpson-Vlach, R. (2013) Corpus Linguistics. U Schmitt N. (ur.), *An introduction to applied linguistics* (pp. 89-105). London: Hodder Education.

Silić, J. & Pranjković, I. (2005) *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.

SketchEngine.eu. (2019) *CQL – Corpus Query Language*. [mrežna stranica] Dostupno na: https://www.sketchengine.eu/documentation/corpus-querying/ (23.06.2019).

Šojat, K., Srebačić, M. i Štefanec, V. (2013). CroDeriV i morfološka raščlamba hrvatskoga glagola. *Suvremena lingvistika, 39* (75), 75-96. Preuzeto s https://hrcak.srce.hr/105505 (05.05.2019)

Spousta, M. (2006). Web as a Corpus. U *Zbornik konference WDS* (pp. 179-184). Dostupno na: https://www.mff.cuni.cz/veda/konference/wds/proc/pdf06/WDS06_132_i3_Spousta.pdf (29.06.2019).

Tadić, M. (1998). Raspon, opseg i sastav korpusa hrvatskoga suvremenog jezika. *Filologija*, (30-31), 337-347.

Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex Libris.

Trenthyer.com. (2019). *What's the difference between Natural Language Processing and Computational Linguistics? – Persephone*. [mrežna stranica] Dostupno na: http://trenthyer.com/comp_ling/whats-the-difference-between-natural-language-processing-and-computational-linguistics/ (17.06.2019).

Tsujii, J. I. (2011). Computational linguistics and natural language processing. *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 52-67).

Popis tablica

Popis grafikona

Analiza distribucije glagolskih vrsta u općem korpusu hrvatskoga jezika

Sažetak

Hrvatske gramatike i preskriptivni priručnici klasificiraju glagole u glagolske vrste prema njihovim morfosintaktičkim karakteristikama. Frekvencijski se podaci o glagolskim vrstama u hrvatskim priručnicima temelje na malom broju glagola. Zaključci poput brojnosti ili produktivnosti nekog razreda doneseni su bez uvida u veliki korpus. Dosadašnje su podjele crpile glagole iz korpusa koji su sastavljeni od književnih djela i kao takvi predstavljaju umjetni oblik hrvatskoga jezika, a gramatike opisuju standardnu inačicu hrvatskoga jezika. U ovom se radu analiziraju glagoli iz korpusa hrWaC koji sadrži 1,4 milijardi pojavnica i oko 90 000 glagola. Analizom glagola iz korpusa hrWaC opisivat će se manje formalni i nestandardni idiom hrvatskoga jezika, odnosno jezik svakodnevice. Oslanjajući se na korpus, rad problematizira glagolsku klasifikaciju u vrste i reevaluira postojeće podatke o glagolskim vrstama u hrvatskim gramatikama. Rad ukratko opisuje glagole i njihove morfološke karakteristike te postojeće klasifikacije. Izlučivanjem glagola iz hrvatskoga mrežnog korpusa hrWaC i njihovim obrađivanjem, odnosno klasifikacijom, rad pruža novi uvid u frekvencijsku distribuciju glagolskih vrsta. Rad ističe probleme i poteškoće koje su se javile pri izradi ovoga modela i preispituje dosadašnje glagolske klasifikacije.

Ključne riječi: korpusna lingvistika, obrada prirodnog jezika, klasifikacija glagola, hrvatske gramatike, hrvatski jezik

Verb class distribution analysis in a general corpus of Croatian language

Summary

Croatian grammar textbooks have a long tradition of classifying verbs based on their morphosyntactic characteristics. These classifications were based on a small number of verbs. Conclusions, such as the frequency or productiveness of a class, were drawn without having the insight into a big corpus. Corpora used in such descriptions were made of literary works and therefore were describing an artificial form of the Croatian language. The corpus used for analyzing verbs in this thesis is hrWaC which contains 1,4 billion tokens and about 90 000. This corpus was selected with the intention of describing and analyzing a less formal and standardized language, i.e. a language more similar to its everyday form. This thesis offers a corpus-based approach to the problem of verb classification and aims to reevaluate the existing concepts and conclusions that are offered in Croatian grammars. The thesis gives a brief introduction to verbs, their morphological characteristics and their classification. By extracting verbs from the Croatian web corpus hrWaC and processing them computationally, the thesis gives an insight into the verb distribution in the Croatian language and points out some difficulties that were encountered during this pilot project of reevaluation of existing verb classifications.

Keywords: corpus linguistics, natural language processing, verb classification, grammar textbooks, Croatian language